METHODS FOR DETECTING AND DIAGNOSING ORAL CANCER

FIELD OF THE INVENTION

The present invention is in the field of diagnostics, detection or research analysis of cancer, and more particularly, oral cancer. More specifically, the present invention is in the field of analysis of the levels of gene expression in specific cancers using microarrays. Even more specifically, embodiments of the present invention are in the field of the identification of biological conditions characterized by alterations of the relative expression levels of various genes.

BACKGROUND

Many cellular events and processes are characterized by altered expression levels of one or more genes. Differences in gene expression correlate with many physiological processes such as cell cycle progression, cell differentiation and cell death. Changes in gene expression patterns also correlate with changes in disease or pharmacological state. For example, the lack of sufficient expression of functional tumor suppressor genes and/or the over expression of oncogene/protooncogenes could lead to tumorgenesis (Marshall, *Cell*, 64: 313-326 (1991); Weinberg, Science, 254: 1138-1146 (1991), incorporated herein by reference in their entireties for all purposes). Thus, changes in the expression levels of particular genes (*e.g.* oncogenes or tumor suppressors) serve as signposts for different physiological, pharmacological and disease states.

Gene expression profiles produce a snapshot that reflects the biological status of the sample, but in many circumstances biological status will reflect more than one characteristic of the sample. For example, when comparing tumor samples from two patients, there will be changes that correlate with differences between the states of the tumors as well as changes that correlate with the different physiological states of the two patients. High-throughput technologies, such as DNA microarrays, have been used to profile and monitor gene expression of hematopoietic tumors (see Alizadeh et al., 2000; Golub et al., 1999 each incorporated by reference in their entireties for all purposes) and solid tumor homogenates and cell lines (see Alon et al., 1999, Perou et al., 2000, Sgroi et al., 1999 each incorporated by reference in their entireties for all purposes). However,

the cell-specific profiling of solid tumor gene expression has been hampered by the inability to procure specific pure cell populations. A need exists to identify genes associated with normal and solid tumor cancerous cell conditions and to further correlate the expression levels of genes as a way of detecting or diagnosing a cancerous condition.

SUMMARY OF THE INVENTION

Embodiments of the present invention are directed to methods of detecting and/or diagnosing cancerous cell conditions, particularly oral cancerous cell conditions, and more particularly solid tumor oral cancerous cell conditions. According to one aspect of the present invention, the expression levels of genes obtained from malignant and non-malignant cell samples are identified and analyzed to provide a gene expression profile for malignant and non-malignant cells. According to one embodiment of the present invention, the expression levels of genes associated with a cell sample are identified and then analyzed and/or compared with known expression levels of genes for malignant and/or normal cells. Based upon similarities of expression levels of genes between the cell sample and malignant cells, a determination can be made as to whether the cell sample is malignant or not or may be predisposed to becoming malignant or not.

Embodiments of the present invention provide for the use of molecular profiling to distinguish normal and malignant oral tissues. Specifically, the present method is useful to diagnose or type oral cancer cells. According to the present invention, markers of malignant cells can be identified and used in diagnostic methods. One aspect of the current invention is directed at identifying genes that are differentially expressed between two biological states as being further correlated with disease, physiological or pharmacological state.

In a first aspect, a method of monitoring expression of one or more genes associated with oral cancer is provided. According to this method, a population of nucleic acids is prepared from a sample of cells obtained from malignant oral tissue. The nucleic acids are then contacted to an array of probes and the relative hybridization of the probes to the nucleic acids is determined. In a second aspect, a method of expression monitoring includes contacting a first array of probes with a first population of nucleic acids derived from at least one cell derived from normal tissue. A second array of probes

is contacted with a second population of nucleic acids derived from at least one cell derived from malignant oral tissue. The relative binding of the probes to the nucleic acids from the first and second populations is then determined to identify at least one probe binding to a gene that is differentially expressed between the first and second populations.

According to the present invention, malignant oral cells can be classified by determining an expression profile of each of a plurality of cells derived from malignant oral tissue, and then classifying the cells in clusters determined by similarity of expression profile.

Embodiments of the present invention are also directed to a method of monitoring differentiation of a malignant oral cell lineage. According to this method, an expression profile of each of a plurality of cells derived from malignant oral tissue at different differentiation stages within the lineage is determined. The cells are classified in clusters determined by similarity of expression profile. The clusters are ordered by similarity of expression profile, and a time course of expression levels for each of the plurality of genes at different stages of differentiation in the malignant oral cell lineage is determined.

Embodiments of the present invention are also directed to a method for identifying differentially expressed transcripts associated with oral cancer. According to this method, an expression profile of each of a plurality of cells derived from malignant oral tissue at different differentiation stages within the lineage is determined. The cells are then classified in clusters determined by similarity of expression profile. The clusters are then ordered by similarity of expression profile, and then a time course of expression levels for each of the plurality of genes at different stages of differentiation in the cell lineage is determined. Differentially expressed transcripts are then identified.

Embodiments of the present invention are still further directed to a method of identifying an oral cancer-associated cell type wherein an expression profile of a plurality of cells derived from malignant oral tissue is determined. The cells are classified in clusters determined by similarity of expression profile, and then the nature and function of a plurality of cells is determined.

Embodiments of the present invention are even still further directed to a method of diagnosing a subject with oral cancer wherein nucleic acids are derived from a sample

of tissue obtained from a subject. The level of expression of at least one gene or marker selected from a group of markers associated with oral cancer is determined. The level of expression of the at least one gene or marker is then compared with the normal level of expression of the marker in a control sample from normal tissue, wherein a difference of degree between the level of expression of the marker in the sample from the subject and the control sample from normal tissue indicates that the subject is afflicted with oral cancer.

A method for monitoring the progression of oral cancer in a subject is also provided wherein the expression of at least one gene or marker selected from a group of markers associated with oral cancer is determined from a sample of tissue taken from a subject at a first point in time. A second sample of tissue is taken at a subsequent point in time and the expression of the at least one gene or marker is determined. The levels of expression are then compared in a manner to monitor the progression of oral cancer.

Still further embodiments of the present invention are directed to a method of assessing the efficacy of a test compound for inhibiting oral cancer in a subject or the efficacy of a therapy for inhibiting oral cancer wherein the ability of a test compound of a therapy to inhibit expression of at least one gene or marker selected from a group of markers associated with oral cancer is determined by comparing the expression levels of the at least on gene or marker with and without the presence of the test compound or the therapy.

Similarly, compounds can be screened for their ability to inhibit oral cancer in a subject by obtaining a sample of cells from the subject, separately maintaining aliquots of the sample in the presence of a plurality of test compounds, comparing expression of at least one gene or marker selected from a group of markers associated with oral cancer in each of the aliquots, and selecting one of the test compounds based on its ability to alter the expression of the gene or marker.

Still, even further embodiments of the present invention are directed to a kit for assessing whether a subject is afflicted with oral cancer. The kit includes reagents and/or an array of probes for assessing expression of at least one gene or marker selected from a group of markers associated with oral cancer.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1A shows amounts of cDNA after two rounds of T7 amplification.

Figure 1B shows percent transcripts detected in normal and tumor tissues.

Figure 2A shows 39 genes whose expression changed in all five samples taken from patients participating in a pilot cancer study.

Figure 2B shows a representative sample of the differentially expressed genes grouped into biological pathways known to be relevant in carcinogenesis

Figure 2C shows candidate genes that are up regulated.

Figure 2D shows candidate genes that are down regulated.

Figure 3 shows a comparison of percent increases for three upregulated genes measured by GeneChip® and Real Time Quantitative PCR data.

Figure 4 shows a comparison of differential collagenase gene expression measured by GeneChip® microarray and RT-QPCR.

Figure 5 shows hierarchical clustering.

Figure 6 shows differentially expressed genes identified by three different methods.

Figure 7 shows differential gene expression using GeneChip® analysis software which revealed that 404 genes are differentially expressed.

DETAILED DESCRIPTION OF CERTAIN PREFERRED EMBODIMENTS

This application relies on, and cites the disclosure of other patent applications and literature references. These documents are hereby incorporated by reference in their entireties for all purposes. The practice of the present invention may employ, unless otherwise indicated, conventional techniques of organic chemistry, polymer technology, molecular biology (including recombinant techniques), cell biology, biochemistry, and immunology, which are within the skill of the art. Such conventional techniques include polymer array synthesis, hybridization, ligation, detection of hybridization using a label. Specific illustrations of suitable techniques can be had by reference to the examples hereinbelow. However, other equivalent conventional procedures can, of course, also be used. Such conventional techniques can be found in standard laboratory manuals such as *Genome Analysis: A Laboratory Manual Series (Vols. I-IV)*, *Using Antibodies: A Laboratory Manual*, *Cells: A Laboratory Manual*, *PCR Primer: A Laboratory Manual*, and *Molecular Cloning: A Laboratory Manual* (all from Cold Spring Harbor Laboratory Press), all of which are herein incorporated in their entirety by reference for all purposes.

This section presents a detailed description of the preferred invention and its application. This description is by way of several exemplary illustrations, in increasing detail and specificity, and of the general methods of this invention. These examples are non-limiting, and related variants that will be apparent to one of skill in the art are intended to be encompassed by the appended claims. Following these examples are descriptions of embodiments of the data gathering steps that accompany the general methods.

Principles of the present invention are directed to the molecular analysis of solid tumors and to providing methods for obtaining information about consistent molecular alterations that advance both the understanding of the basic biology of tumors as well as the clinically relevant aspects of the molecular epidemiology of oral cancer. In one aspect, the present invention incorporates the use of laser capture microdissection-derived RNA to be used on microarrays and that array hybridization coupled with hierarchical and non-hierarchical analysis methods provide powerful approaches for identifying candidate genes and molecular profiling associated with oral cancer.

Markers according to the present invention may include any nucleic acid sequence or molecule or corresponding polypeptide encoded by the nucleic acid sequence or molecule which demonstrates altered expression (i.e., higher or lower expression) in oral cancer samples relative to normal samples (i.e., non-oral cancer samples).

Nucleic acids according to the present invention may include any polymer or oligomer of pyrimidine and purine bases, preferably cytosine, thymine, and uracil, and adenine and guanine, respectively. *See* Albert L. Lehninger, PRINCIPLES OF BIOCHEMISTRY, at 793-800 (Worth Pub. 1982). Indeed, the present invention contemplates any deoxyribonucleotide, ribonucleotide or peptide nucleic acid component, and any chemical variants thereof, such as methylated, hydroxymethylated or glucosylated forms of these bases, and the like. The polymers or oligomers may be heterogeneous or homogeneous in composition, and may be isolated from naturally-occurring sources or may be artificially or synthetically produced. In addition, the nucleic acids may be DNA or RNA, or a mixture thereof, and may exist permanently or transitionally in single-stranded or double-stranded form, including homoduplex,

heteroduplex, and hybrid states. Oligonucleotide and polynucleotide are included in this definition and relate to two or more nucleic acids in a polynucleotide.

Peptide: A polymer in which the monomers are alpha amino acids and which are joined together through amide bonds, alternatively referred to as a polypeptide and/or protein. In the context of this specification it should be appreciated that the amino acids may be, for example, the L-optical isomer or the D-optical isomer. Peptides are often two or more amino acid monomers long, and often 4 or more amino acids long, often 5 or more amino acids long, often 10 or more amino acids long, often 15 or more amino acids long, and often 20 or more amino acid monomers long, for example. Standard abbreviations for amino acids are used (e.g., P for proline). These abbreviations are included in Stryer, Biochemistry, Third Ed., 1988, which is incorporated herein by reference in its entirety for all purposes.

Array: An array comprises a solid support with peptide or nucleic acid probes attached to said support. Arrays typically comprise a plurality of different nucleic acid or peptide probes that are coupled to a surface of a substrate in different, known locations. These arrays, also described as "microarrays" or colloquially "chips" have been generally described in the art, for example, U.S. Pat. Nos. 5,143,854, 5,445,934, 5,744,305, 5,677,195, 6,040,193, 5,424,186 and Fodor et al., Science, 251:767-777 (1991). Each of which is incorporated by reference in its entirety for all purposes. These arrays may generally be produced using mechanical synthesis methods or light directed synthesis methods which incorporate a combination of photolithographic methods and solid phase synthesis methods. Techniques for the synthesis of these arrays using mechanical synthesis methods are described in, e.g., U.S. Pat. No. 5,384,261, and 6,040,193 which are incorporated herein by reference in their entirety for all purposes. Although a planar array surface is preferred, the array may be fabricated on a surface of virtually any shape or even a multiplicity of surfaces. Arrays may be peptides or nucleic acids on beads, gels, polymeric surfaces, fibers such as fiber optics, glass or any other appropriate substrate, see U.S. Patent Nos. 5,770,358, 5,789,162, 5,708,153, 6,040,193 and 5,800,992, which are hereby incorporated by reference in their entirety for all purposes. Arrays may be packaged in such a manner as to allow for diagnostics or other manipulation of in an all inclusive device, see for example, US Patent Nos. 5,856,174 and 5,922,591 incorporated

Express Mail Label No. EL 977 950 705 US

in their entirety by reference for all purposes. See also attorney docket number 3233.1 for additional information concerning arrays, their manufacture, and their characteristics. It is hereby incorporated by reference in its entirety for all purposes.

Gene expression monitoring is a useful way to distinguish between cells that express different phenotypes. For example, cells that are derived from different organs, have different ages, or different physiological states. In a preferred embodiment, gene expression monitoring can distinguish between cancer cells and normal cells, or different types of cancer cells.

Expression profile: One measurement of cellular constituents that is particularly useful in the present invention is the expression profile. As used herein, an "expression profile" comprises measurement of the relative abundance of a plurality of cellular constituents. Such measurements may include RNA or protein abundances or activity levels. An expression profile involves providing a pool of target nucleic acid molecules or polypeptides, hybridizing the pool to an array of probes immobilized on predetermined regions of a surface, and quantifying the hybridized nucleic acid molecules or proteins. The expression profile can be a measurement, for example, of the transcriptional state or the translational state of the cell. See U.S. Patent Nos. 6,040,138, 6,013,449 and 5,800,992, which are hereby incorporated by reference in their entireties for all purposes.

Transcriptional state: The transcriptional state of a sample includes the identities and relative abundances of the RNA species, especially mRNAs present in the sample. Preferably, a substantial fraction of all constituent RNA species in the sample are measured, but at least a sufficient fraction is measured to characterize the state of the sample. The transcriptional state is the currently preferred aspect of the biological state measured in this invention. It can be conveniently determined by measuring transcript abundances by any of several existing gene expression technologies.

Translational state: Translational state includes the identities and relative abundances of the constituent protein species in the sample. As is known to those of skill in the art, the transcriptional state and translational state are related.

The gene expression monitoring system, in a preferred embodiment, may comprise a nucleic acid probe array (such as those described above), membrane blot (such as used in hybridization analysis such as Northern, Southern, dot, and the like),

microwells, sample tubes, gels, beads or fibers (or any solid support comprising bound nucleic acids). *See* U.S. Patent Nos. 5,770,722, 5,874,219, 5,744,305, 5,677,195 and 5,445,934, 5,800,992 which are expressly incorporated herein by reference in their entireties for all purposes.

The gene expression monitoring system according to the present invention may be used to facilitate a comparative analysis of expression in different cells or tissues, different subpopulations of the same cells or tissues, different physiological states of the same cells or tissue, different developmental stages of the same cells or tissue, or different cell populations of the same tissue.

Differentially expressed: The term differentially expressed as used herein means that a measurement of a cellular constituent varies in two samples. The cellular constituent can be either upregulated in the experiment relative to the reference or downregulated in the experiment relative to the reference. Differential gene expression can also be used to distinguish between cell types or nucleic acids. See U.S. Patent No. 5,800,992.

One of skill in the art will appreciate that it is desirable to have nucleic acid samples containing target nucleic acid sequences that reflect the transcripts of interest. Therefore, suitable nucleic acid samples may contain transcripts of interest. Suitable nucleic acid samples, however, may contain nucleic acids derived from the transcripts of interest. As used herein, a nucleic acid derived from a transcript refers to a nucleic acid for whose synthesis the mRNA transcript or a subsequence thereof has ultimately served as a template. Thus, a cDNA reverse transcribed from a transcript, an RNA transcribed from that cDNA, a DNA amplified from the cDNA, an RNA transcribed from the amplified DNA, etc., are all derived from the transcript and detection of such derived products is indicative of the presence and/or abundance of the original transcript in a sample. Thus, suitable samples include, but are not limited to, transcripts of the gene or genes, cDNA reverse transcribed from the transcript, cRNA transcribed from the cDNA, DNA amplified from the genes, RNA transcribed from amplified DNA, and the like.

Transcripts, as used herein, may include, but are not limited to pre-mRNA nascent transcript(s), transcript processing intermediates, mature mRNA(s) and degradation products. It is not necessary to monitor all types of transcripts to practice this invention.

For example, one may choose to practice the invention to measure the mature mRNA levels only.

In one embodiment, a sample is a homogenate of cells (e.g., oral cells and/or blood cells), tissues or other biological samples. Preferably, such sample is a nucleic acid preparation, e.g., a total RNA preparation of a biological sample. More preferably in some embodiments, such a nucleic acid sample is the total mRNA isolated from a biological sample. Those of skill in the art will appreciate that the total mRNA prepared with most methods includes not only the mature mRNA, but also the RNA processing intermediates and nascent pre-mRNA transcripts. For example, total mRNA purified with a poly (T) column contains RNA molecules with poly (A) tails. Those poly A+ RNA molecules could be mature mRNA, RNA processing intermediates, nascent transcripts or degradation intermediates.

Biological samples may be of any biological tissue or fluid or cells. Frequently the sample will be a "clinical sample" which is a sample derived from a patient. Clinical samples provide rich sources of information regarding the various states of genetic network or gene expression. Some embodiments of the invention are employed to detect mutations and to identify the function of mutations. Such embodiments have extensive applications in clinical diagnostics and clinical studies. Typical clinical samples include, but are not limited to, sputum, blood, blood cells (e.g., white cells), tissue or fine needle biopsy samples, urine, peritoneal fluid, and pleural fluid, or cells therefrom. Biological samples may also include sections of tissues such as frozen sections taken for histological purposes.

Another typical source of biological samples are cell cultures where gene expression states can be manipulated to explore the relationship among genes. In one aspect of the invention, methods are provided to generate biological samples reflecting a wide variety of states of the genetic network.

In a preferred embodiment, the level of expression of a marker for oral cancer is assessed by detecting the presence of a nucleic acid corresponding to the marker in the sample. In another preferred embodiment, the level of expression of a marker for oral cancer is assessed by detecting the presence of a protein corresponding to the marker in the sample. In a preferred aspect, the presence of the protein is detected using a reagent

which specifically binds to the protein, e.g., an antibody, an antibody derivative, and/or an antibody fragment.

Detection involves contacting a sample with a compound or an agent capable of detecting a marker associated with oral cancer such that the presence of the marker is detected in the biological sample. A preferred agent for detecting marker RNA is a labeled or labelable nucleic acid probe capable of hybridizing to marker RNA. The nucleic acid probe can be, for example, complementary to any of the nucleic acid markers of oral cancer disclosed herein, or a portion thereof, such as an oligonucleotide which specifically hybridizes marker RNA.

A preferred agent for detecting a marker protein is a labeled or labelable antibody capable of binding to the marker protein. Antibodies can be polyclonal, or more preferably, monoclonal. An intact antibody, antibody derivative, or a fragment thereof (e.g., Fab or $F(ab')_2$) can be used. The term "labeled or labelable", with regard to the probe or antibody, is intended to encompass direct labeling of the probe or antibody by coupling (i.e., physically linking) a detectable substance to the probe or antibody, as well as indirect labeling of the probe or antibody by reactivity with another reagent that is directly labeled. Examples of indirect labeling include detection of a primary antibody using a fluorescently labeled secondary antibody and end-labeling of a DNA probe with biotin such that it can be detected with fluorescently labeled streptavidin.

The detection methods described herein can be used to detect marker RNA or marker protein in a biological sample *in vitro* as well as *in vivo*. *In vitro* techniques for detection of marker RNA include, but are not limited to, Northern hybridizations and *in situ* hybridizations. *In vitro* techniques for detection of marker protein include, but are not limited to, enzyme linked immunosorbent assays (ELISAs), Western blots, immunoprecipitations, and immunofluorescence assays. Alternatively, marker protein can be detected *in vivo* in a subject by introducing into the subject a labeled antibody against the marker protein. For example, the antibody can be labeled with a radioactive marker whose presence and location in a subject can be detected by standard imaging techniques.

One of skill in the art would appreciate that it is desirable to inhibit or destroy RNase present in homogenates before homogenates can be used for hybridization.

Methods of inhibiting or destroying nucleases are well known in the art. In some preferred embodiments, cells or tissues are homogenized in the presence of chaotropic agents to inhibit nuclease. In some other embodiments, RNases are inhibited or destroyed by heat treatment followed by proteinase treatment.

Methods of isolating total mRNA are also well known to those of skill in the art. For example, methods of isolation and purification of nucleic acids are described in detail in Chapter 3 of Laboratory Techniques in Biochemistry and Molecular Biology: Hybridization With Nucleic Acid Probes, Part I. Theory and Nucleic Acid Preparation, P. Tijssen, ed. Elsevier, N.Y. (1993) and Chapter 3 of Laboratory Techniques in Biochemistry and Molecular Biology: Hybridization With Nucleic Acid Probes, Part I. Theory and Nucleic Acid Preparation, P. Tijssen, ed. Elsevier, N.Y. (1993)).

In a preferred embodiment, total RNA is isolated from a given sample using, for example, an acid guanidinium-phenol-chloroform extraction method followed by polyA$^+$ mRNA isolation by oligo dT column chromatography or by using (dT)n magnetic beads (see, e.g., Sambrook et al., Molecular Cloning: A Laboratory Manual (2nd ed.), Vols. 1-3, Cold Spring Harbor Laboratory, (1989), or Current Protocols in Molecular Biology, F. Ausubel et al., ed. Greene Publishing and Wiley-Interscience, New York (1987) each hereby incorporated by reference in their entireties for all purposes). See also PCT/US99/25200 for complexity management and other sample preparation techniques, which is hereby incorporated by reference in its entirety for all purposes.

Frequently, it is desirable to amplify the nucleic acid sample prior to hybridization. One of skill in the art will appreciate that methods of amplifying nucleic acids are well known in the art and that whatever amplification method is used, if a quantitative result is desired, care must be taken to use a method that maintains or controls for the relative frequencies of the amplified nucleic acids to achieve quantitative amplification.

Methods of "quantitative" amplification are well known to those of skill in the art. For example, quantitative PCR involves simultaneously co-amplifying a known quantity of a control sequence using the same primers. This provides an internal standard that may be used to calibrate the PCR reaction. A high density array may then be performed

which includes probes specific to the internal standard for quantification of the amplified nucleic acid.

Other suitable amplification methods include, but are not limited to polymerase chain reaction (PCR) (Innis, et al., PCR Protocols. A guide to Methods and Application. Academic Press, Inc. San Diego, (1990)), ligase chain reaction (LCR) (see Wu and Wallace, Genomics, 4: 560 (1989), Landegren, et al., Science, 241: 1077 (1988) and Barringer, et al., Gene, 89: 117 (1990)), transcription amplification (Kwoh, et al., Proc. Natl. Acad. Sci. USA, 86: 1173 (1989)), and self-sustained sequence replication (Guatelli, et al., Proc. Nat. Acad. Sci. USA, 87: 1874 (1990)).

Cell lysates or tissue homogenates often contain a number of inhibitors of polymerase activity. Therefore, the skilled practitioner typically incorporates preliminary steps to isolate total RNA or mRNA for subsequent use as an amplification template. One tube mRNA capture methods may be used to prepare poly(A)+ RNA samples suitable for immediate RT-PCR in the same tube (Boehringer Mannheim). The captured mRNA can be directly subjected to RT-PCR by adding a reverse transcription mix and, subsequently, a PCR mix.

In a particularly preferred embodiment, the sample mRNA is reverse transcribed with a reverse transcriptase and a primer consisting of oligo dT and a sequence encoding the phage T7 promoter to provide single stranded DNA template. The second DNA strand is polymerized using a DNA polymerase. After synthesis of double-stranded cDNA, T7 RNA polymerase is added and RNA is transcribed from the cDNA template. Successive rounds of transcription from each single cDNA template results in amplified RNA. Methods of in vitro polymerization are well known to those of skill in the art (see, e.g., Sambrook, supra).

It will be appreciated by one of skill in the art that the direct transcription method described above provides an antisense RNA (aRNA) pool. Where aRNA is used as the target nucleic acid, the oligonucleotide probes provided in the array are chosen to be complementary to subsequences of the antisense nucleic acids. Conversely, where the target nucleic acid pool is a pool of sense nucleic acids, the oligonucleotide probes are selected to be complementary to subsequences of the sense nucleic acids. Finally, where

the nucleic acid pool is double stranded, the probes may be of either sense as the target nucleic acids include both sense and antisense strands.

The protocols cited above include methods of generating pools of either sense or antisense nucleic acids. Indeed, one approach can be used to generate either sense or antisense nucleic acids as desired. For example, the cDNA can be directionally cloned into a vector (e.g., Stratagene's p Bluscript II KS (+) phagemid) such that it is flanked by the T3 and T7 promoters. In vitro transcription with the T3 polymerase will produce RNA of one sense (the sense depending on the orientation of the insert), while in vitro transcription with the T7 polymerase will produce RNA having the opposite sense. Other suitable cloning systems include phage lambda vectors designed for Cre-loxP plasmid subcloning (see e.g., Palazzolo et al., Gene, 88: 25-36 (1990)).

Other analysis methods that can be used in the present invention include electrochemical denaturation of double stranded nucleic acids, U.S. Pat. No. 6,045,996 and 6,033,850, the use of multiple arrays (arrays of arrays), U.S. Pat. No. 5,874,219, the use of scanners to read the arrays, U.S. Pat. Nos. 5,631,734; 5,744,305; 5,981,956 and 6,025,601, methods for mixing fluids, U.S. Pat. No. 6,050,719, integrated device for reactions, U.S. Pat. No. 6,043,080, integrated nucleic acid diagnostic device, U.S. Pat. No. 5,922,591, and nucleic acid affinity columns, U.S. Pat. No. 6,013,440. All of the above patents are hereby incorporated by reference in their entireties for all purposes.

Laser dissection microscopy is one method that can be used in the present invention. That technique is be shown in provisional application 60/182,452 (attorney docket number 3294) which is hereby incorporated by reference in its entirety for all purposes. Other techniques include L. Zhang et al., Science 276, 1268 (1997), Mahadevappa, M. & Warrington, J. A. *Nat. Biotechnol.* **17**, 1134-1136 (1999) and Luo, L. et al. *Nature Med.* **5**, 117-122 (1999) which are all hereby incorporated by reference in their entireties for all purposes.

In a preferred embodiment, the invention provides methods of assessing the efficacy of test compounds and compositions for treating oral cancer. The methods entail identifying candidate or test compounds or agents (e.g., peptides, peptidomimetics, small molecules or other drugs) which have an inhibitory effect on oral cancer. Candidate or test compounds or agents which have an inhibitory effect on oral cancer are identified in

assays that employ oral cancer cells, such as an expression assay entailing direct or indirect measurement of the expression of an oral cancer marker (e.g., a nucleic acid marker or a protein marker). For example, modulators of expression of oral cancer markers can be identified in a method in which a cell is contacted with a candidate compound and the expression of oral cancer markers (e.g., nucleic acid markers and/or protein markers) in the cell is determined. The level of expression of oral cancer markers in the presence of the candidate compound is compared to the level of expression of oral cancer markers in the absence of the candidate compound. The candidate compound can then be identified as a modulator of expression of oral cancer based on this comparison.

The invention also encompasses kits for assessing whether a subject is afflicted with oral cancer, as well as kits for assessing the presence of oral cancer cells. The kit may comprise a labeled compound or agent capable of detecting oral cancer markers (e.g., nucleic acid markers and/or protein markers) in a biological sample, a means for determining the amount of oral cancer markers in the sample, and a means for comparing the amount of oral cancer markers in the sample with a standard. The compound or agent can be packaged in a suitable container. The kit can further comprise instructions for using the kit to detect oral cancer markers.

Those skilled in the art will recognize that in a preferred embodiment, the expression profiles from the reference samples will be input to a database. A relational database is preferred and can be used, but one of skill in the art will recognize that other databases could be used. A relational database is a set of tables containing data fitted into predefined categories. Each table, or relation, contains one or more data categories in columns. Each row contains a unique instance of data for the categories defined by the columns. For example, a typical database for the invention would include a table that describes a sample with columns for age, gender, reproductive status, expression profile and so forth. Another table would describe a disease: symptoms, level, sample identification, expression profile and so forth. See U.S. Ser. No. 09/354,935, which is hereby incorporated by reference in its entirety for all purposes.

In one embodiment the invention matches the experimental sample to a database of reference samples. The database is assembled with a plurality of different samples to be used as reference samples. An individual reference sample in one embodiment will be

obtained from a patient during a visit to a medical professional. The sample could be, for example, a tissue, blood, urine, feces or saliva sample. Information about the physiological, disease and/or pharmacological status of the sample will also be obtained through any method available. This may include, but is not limited to, expression profile analysis, clinical analysis, medical history and/or patient interview. For example, the patient could be interviewed to determine age, sex, ethnic origin, symptoms or past diagnosis of disease, and the identity of any therapies the patient is currently undergoing. A plurality of these reference samples will be taken. A single individual may contribute a single reference sample or more than one sample over time. One skilled in the art will recognize that confidence levels in predictions based on comparison to a database increase as the number of reference samples in the database increases. One skilled in the art will also recognize that some of the indicators of status will be determined by less precise means, for example information obtained from a patient interview is limited by the subjective interpretation of the patient. Additionally, a patient may lie about age or lack sufficient information to provide accurate information about ethnic or other information. Descriptions of the severity of disease symptoms is a particularly subjective and unreliable indicator of disease status.

The database is organized into groups of reference samples. Each reference sample contains information about physiological, pharmacological and/or disease status. In one aspect the database is a relational database with data organized in three data tables, one where the samples are grouped primarily by physiological status, one where the samples are grouped primarily by disease status, and one where the samples are grouped primarily by pharmacological status. Within each table the samples can be further grouped according to the two remaining categories. For example, the physiological status table could be further categorized according to disease and pharmacological status.

As will be appreciated by one of skill in the art, the present invention may be embodied as a method, data processing system or program products. Accordingly, the present invention may take the form of data analysis systems, methods, analysis software and etc. Software written according to the present invention is to be stored in some form of computer readable medium, such as memory, hard-drive, DVD ROM or CD ROM, or transmitted over a network, and executed by a processor. The present invention also

provides a computer system for analyzing physiological states, levels of disease states and or therapeutic efficacy. The computer system comprises a processor, and memory coupled to said processor which encodes one or more programs. The programs encoded in memory cause the processor to perform the steps of the above methods wherein the expression profiles and information about physiological, pharmacological and disease states are received by the computer system as input.

U.S. Patent No. 5,733,729 illustrates an example of a computer system that may be used to execute the software of an embodiment of the invention. This patent shows a computer system that includes a display, screen, cabinet, keyboard, and mouse. The mouse may have one or more buttons for interacting with a graphic user interface. The cabinet preferably houses a CD-ROM or DVD-ROM drive, system memory and a hard drive which may be utilized to store and retrieve software programs incorporating computer code that implements the invention, data for use with the invention and the like. Although a CD is shown as an exemplary computer readable medium, other computer readable storage media including a floppy disk, a tape, a flash memory, a system memory, and a hard drive may be utilized. Additionally, a data signal embodied in a carrier wave (*e.g.,* in a network including the internet) may be the computer readable storage medium.

The patent also shows a system block diagram of a computer system used to execute the software of an embodiment of the invention. The computer system includes a monitor, a keyboard, and a mouse. The computer system further includes subsystems such as a central processor, a system memory, a fixed storage (*e.g.*, a hard drive), a removable storage (*e.g.*, CD-ROM), a display adapter, a sound card, speakers, and a network interface. Other computer systems suitable for use with the invention may include additional or fewer subsystems. For example, another computer system may include more than one processor or a cache memory. Computer systems suitable for use with the invention may also be embedded in a measurement instrument. The embedded systems may control the operation of, for example, a GeneChip® Probe array scanner (also called a GeneArray™ scanner sold by Agilent corporation, Palo Alto Ca.) as well as executing computer codes of the invention.

Computer methods can be used to measure the variables and to match samples to eliminate gene expression differences that are a result of differences that are not of interest. For example, a plurality of values can be input into computer code for one or more physiological, pharmacological and/or disease states. The computer code can thereafter measure the differences or similarities between the values to eliminate changes not attributable to a value of interest. Examples of computer programs and databases that can be used for this purpose are shown in U.S.S.Nos. 09/354,935, 08/828,952, 09/341,302, 09/397,494, 60/220587, and 60/220645, which are hereby incorporated by reference in their entireties for all purposes.

Computer software to analyze data generated by microarrays is commercially available from Affymetrix Inc. (Santa Clara) as well as other companies. Affymetrix Inc. distributes GeneChip®, now known as MicroArray suite, LIMS, Microdb, Jaguar, DMT, and other software. Other databases can be constructed using the standard database tools available from Microsoft (e.g., Excel and Access).

In one aspect of the invention, microarrays will be used to measure expression profiles. Microarrays are particularly well suited because of the reproducibility between different experiments. DNA microarrays provide one method for the simultaneous measurement of the expression levels of large numbers of genes. Each array consists of a reproducible pattern of thousands of different DNAs attached to a solid support. Labeled RNA or DNA is hybridized to complementary probes on the array and then detected by laser scanning. Hybridization intensities for each probe on the array are determined and converted to a quantitative read-out of relative gene expression levels. The data can be further analyzed to identify expression patterns and variations that correlate with the biological state of the sample. (See U.S. Patent Nos. 6,040,138, and 5,800,992 which are incorporated herein by reference in their entireties for all purposes.)

High-density oligonucleotide arrays are particularly useful for monitoring the gene expression pattern of a sample. In one approach, total mRNA isolated from the sample is converted to labeled cRNA and then hybridized to an array such as a GeneChip® oligonucleotide array. Each sample is hybridized to a separate array. Relative transcript levels are calculated by reference to appropriate controls present on the array and in the sample.

EXAMPLE I

Preparing Nucleic Acid Samples

Tissue samples were obtained from 5 patients as described in Table 1 below. Each of patients 1-5 has a history of smoking and alcohol consumption, which are the major etiological causes of oral cancer. Each patient exhibited squamous cell carcinoma (SCC): patient 1 was moderately differentiated, patients 2 and 3 were well differentiated, patient 4 was moderately differentiated and patient 5 was moderate to poorly differentiated. Normal tissue is designated as "A" and tumor tissue is designated as "B". the term "ppd" means packs per day and "cig" means either a cigarette or a cigar.

Table 1

Identifier	Gender	Age	Smoking History	Alcohol Consumption
1A, 1B	F	80	Not Known	Not Known
2A, 2B	M	61	2ppd/15 yrs	2 shots/day
3A, 3B	M	68	1-2 ppd/40 yrs	3 beers/day
4A, 4B	M	75	1-2 cig/day	2 drinks/day
5A, 5B	F	60	40 pack/yr	Heavy

Normal and tumor cells from a solid tumor site from within the oral cavity were obtained using laser capture microdissection as described in provisional application 60/182,452 (attorney docket number 3294) which is hereby incorporated by reference in its entirety for all purposes. According to that method, biopsies were taken and snap frozen. The biopsies were sectioned at 5 microns and mounted on slides. They were then stained with hematoxylin and eosin. Laser capture microdissection was then used to procure malignant and normal keratinocytes. Laser capture microdissection, RNA isolation, IVT, 3 rounds of T7 RNA polymerase linear amplification, and probe biotinylation were carried out according to the methods of Alevizos et al., submitted, (2000) and Ohyama et al., Biotechniques 29, 530-6 (2000), each of which are hereby incorporated by reference in their entireties for all purposes. Basically, RNA was extracted and then cDNA synthesis was carried out using Superscript (Life Technologies). cRNA synthesis and labeling was carried out using Ampliscribe

(Epicenter technology) and BioArray High Yield RNA Transcript Labeling System (Enzo).

The quality and quantity of isolated RNA was examined by reverse transcription polymerase chain reaction (RT-PCR) of five cellular maintenance gene transcripts of high to low abundance (glyceraldehyde-3-phosphate dehydrogenase, tubulin-α, β-actin, ribosomal protein S9, and ubiquitin C) (Ohyama et al., 2000). The quantity of isolated RNA was also assessed with RiboGreen RNA Quantitation Reagent and kit (Molecular Probes, Eugene, OR) using spectrofluorometry (Bio-Rad, Hercules, CA). Only those samples exhibiting PCR products for all five cellular maintenance genes were used for subsequent analysis. The biotinylated cRNA from the ten samples (five normal and five cancer) were further used to hybridize the Affymetrix Test-1 probe arrays to determine cRNA quality and integrity. The arrays contain probes representing a handful of maintenance genes and a number of controls (Ohyama et al., 2000). Analysis of the arrays confirmed the RT-PCR findings. cRNA linearly amplified from human oral cancer tissue produced no nonspecific or unusual hybridization patterns, and transcripts corresponding to the maintenance genes were detected. The 5' region of the RNA was degraded, but sufficient 3' transcript was intact to proceed with hybridization using the HuGeneFL probe arrays. In addition, probes synthesized on the arrays are biased to the last 600 bp in the 3' region of the transcripts. Yields of cDNA resulting from the LCM, RNA isolation, and after two rounds of T7 amplification are shown in Figure 1A. Linear amplification of total RNA began with ~100 ng of total RNA. As shown in Figure 1A, the amount of double stranded cDNA (ds-cDNA) after two rounds of T7 amplification is dependent on the quality of the LCM-generated RNA from the normal and tumor tissues.

Fig. 1B summarizes the hybridization outcome of the five paired cases of oral cancers. The percent transcript detected ranged from 26 to 40 percent, indicating satisfactory quality and representation of the harvested RNA. Note that the difference between the normal and cancer samples from each patient is very similar, indicating little variability among each pair, suggesting that the quality of the RNA isolated from the normal and tumor epithelium is similar.

EXAMPLE II

Analysis of Nucleic Acid Samples Using MicroArrays

The cRNA was fragmented as described by Wodicka et al. (1997) and then hybridized to Affymetrix probe arrays such as GeneChip Test 1, Human U95A and HuGeneFL probe arrays. Hybridization was carried out for a time period of between about 12 to 16 hours. All array washing, staining and scanning were carried out as described in the Gene Expression Manual (Affymetrix, Inc. 1999 hereby incorporated by reference in its entirety for all purposes). The Affymetrix arrays include probe sets consisting of oligonucleotides 25 bases in length. Probes are complementary to the published sequences (GeneBank) as previously described (Lockhart et al., 1996). The sensitivity and reproducibility of the GeneChip® probe arrays is such that RNAs present at a frequency of 1:100,000 are unambiguously detected, and detection is quantitative over more than three orders of magnitude (Redfern et al., 2000; Warrington et al., 2000). In this set of experiments with oral cancer samples, the bacterial transcript (BioB), spiked before the hybridization at concentration of 1.5 pM. This concentration, which corresponds to three copies per cell, based on the assumption that there are 300,000 transcripts per cell with an average transcript length of 1 kb, was defined as present in nine out of ten experiments (Lockhart et al., 1996). Array controls, and performance with respect to specificity and sensitivity are the same as those previously described (Lockhart et al., 1996; Mahadevappa & Warrington, 1999; Wodicka et al., 1997). Information regarding the genes represented on the arrays used in this experiment can be found at www.netaffx.com.

EXAMPLE III

Software Analysis of Data

The data obtained from the microarrays was analyzed using various methods and software commercially available and known to those skilled in the art. These methods and software include T-test, GeneChip® software available from Affymetrix to perform a comparison analysis; GeneCluster SOM software to perform a cluster analysis, identify genes and develop characteristics of gene expression profiles; and Matlab software to

identify genes that are differentiating and to identify gene classes. Additional software that can be used to analyze chip data includes GenExplore and PCA.

For GeneCluster analysis and the computation of self organizing maps (SOM), gene expression levels and geometry of nodes were input into the GeneCluster software. Before the computation of the SOM, two preprocessing steps took place. First, a filter was applied to exclude genes that did not change significantly across the pairs. Genes were eliminated if they did not show a relative change of x=2 and an absolute change of y=35, (x, y)=(2, 35). Second, normalization of expression levels across experiments was carried out, thus emphasizing the expression pattern rather than the absolute expression values. Data was normalized using GeneChip® software. A description of the normalization procedure can be found on pp. A5-14, GeneChip® Expression Analysis Technical Manual, (Tamayo et al., 1999).

Differential gene expression using GeneChip® analysis software revealed that 404 probe sets changed in the majority of the cases (3/5) (set forth in Figure 7). Among the 404, 211 were increased in tumor samples and 193 were decreased in tumor samples, compared to normal samples. As shown in Figure 2A, 39 probe sets used allowed the detection of changes in gene expression in all five cases. Sixteen genes showed increased expression in tumor samples and 23 genes showed decreased expression in tumor samples, compared to normal samples. Figure 2B is a list of differentially expressed genes grouped into biological pathways known to be relevant in carcinogenesis. Figure 2C is a list of differentially expressed genes which are up regulated in tumor samples. Figure 2D is a list of differentially expressed genes which are down regulated in tumor samples.

The data presented in Figs. 2A and 2B shows that many known genes involved in neoplasia are differentially expressed in the five paired cases of oral cancer. Thus, these genes are markers associated with oral cancer. Further, the data indicates that the expression of members of known biological pathways are altered during oral carcinogenesis. These include genes which regulate metastatic and invasion pathways, transcription factors, oncogenes and tumor suppressor genes, and differentiation markers (Figure 2B). Of particular importance are the differentially expressed genes that are not yet fully functionally characterized or genes that have not been studied by classic

methods in head and neck/ oral carcinogenesis. One such example is neuromedin U (Nmu), which is downregulated in five out of five tumors (Szekeres et al., 2000). Nmu is a poorly understood protein that manifests potent contractile activities on smooth muscle cells. Recently, two G-protein coupled receptors (NMU1 and NMU2) have been identified to interact with Nmu with nanomolar potency (Fujii et al., 2000; Raddatz et al., 2000). Surprisingly, Nmu is relevant in the development of oral malignancy and can function as a marker for carcinogenesis.

In order to validate the expression level data, three metastatic pathway genes whose expression are consistently altered in the five paired cases of oral cancer were selected. Real-time quantitative PCR (RT-QPCR) in conjunction with the TaqMan specific probe system or SYBR® Green system were used to validate the expression levels of interstitial collagenase (a member of the MMP's involved in metastasis), urokinase plasminogen activator (UPA, associated with metastasis) and cathepsin L (a member of the serine proteases). The cDNA product of the reverse transcription reaction was used as the template for the RT-QPCR reaction. For the RT-QPCR reaction, the iCycler IQ™ Real Time PCR detection system (Bio-Rad, Hercules, CA) was used with TaqMan specific probes and primers for Cathepsin, and SYBR® Green buffer and reagents (Perkin Elmer/Applied Biosystems Foster City, CA, USA) for Urokinase Plasminogen Activator and Collagenase I (Heid et al., 1996).

For designing the specific primers and probes, PE/ABD Primer Express software as well as MacVector were used. Primer sequences used were:

Collagenase forward: 5'-ACACGGAACCCCAAGGACA-3'

Collagenase Reverse: 5'-GTTTTGTTGCCGGTGGTTTT-3'

UPA forward: 5'-GCACCATCAAACAAACCCCCTTAC-3'

UPA reverse: 5'-CAGACAGAAAAACCCCTGCCTG-3'

Cathepsin L forward: 5'-CAGTGTGGTTCTTGTTGGGCT-3'

Cathepsin L reverse: 5'-CTTGAGGCCCAGAGCAGTCTA-3'

The final PCR products were run on 2% minigel to ensure single product amplification during the PCR assay.

Comparison of the microarray and RT-QPCR data as shown in Figure 3 revealed that they approximate each other. The slight observed discrepancy in the precise

quantitation of the GeneChip® and the RT-QPCR was due to the fact that a minute amount (ng) of LCM-generated total RNA was used for amplification followed by biotinylation and hybridization to the GeneChip® microarrays. Using the same LCM-generated total RNA, the GeneChip® data of three metastatic tumor genes was validated by real-time quantitative PCR (RT-QPCR). These two independent approaches yielded data which indicated a similar trend (Figure 4). Both methods indicate that genes were upregulated from undetectable levels in the control to moderate abundance in the tumor cells. Similar results of GeneChip® versus RT-QPCR correlation were previously used by Welsh et al. to validate candidates identified in an ovarian cancer study (Welch et al., 2001). The RT-QPCR data confirmed the upregulation and downregulation of selected candidates. Therefore, while there is discrepancy in the precise quantitation of GeneChip® and RT-QPCR data of each sample, the overall trend and correlation are similar. The array data produces information about relative abundance that is accurate to within 1.5 to 2 fold (Lockhart et al., 1996; Redfern et al., 2000) providing information that allows binning of the transcript levels by low, low-medium, medium, medium-high or high abundance (Warrington et al., 2000a; Warrington et al., 2000b; Lockhart et al., 1996; Redfern et al., 2000).

The actual comparative data for collagenase is graphically depicted in Figure 4. As presented in Figure 3, similar data were obtained for UPA & cathepsin L. Other high and low genes including Neuromedin U, GST, cytochrome P450, ALDH-9, ALDH-10 and Wilm's tumor-related protein have also been validated.

The microarray data can be analyzed by pattern recognition (clustering) software to aid in deriving lists of genes that distinguish and characterize disease versus normal biopsies, thus shedding light on molecular genetic profiles and ultimately the mechanism of the disease under study. Techniques used for conducting hierarchical clustering analyses include self-organizing maps (SOM), Bayesian, hierarchical, and k-means. SOM was selected because of advantages in initial exploration of the data allowing the operator to impose partial structure on the clusters (Tamayo et al., 1999). Other advantages of SOM include good computational properties, computational speed and ease of implementation. SOM analysis was applied to the microarray data on the five paired cases of oral cancers. The clusters graphically represent gene expression patterns across

all ten samples (normal and tumor), each cluster differing in gene number and grouping. This method provides a candidate set of genes whose differing expression activity can be used to distinguish normal and tumor cell behavior.

By SOM analysis using GeneCluster software, 178 transcripts were found to be differentially expressed between tumor and normal tissues. An important observation is that many of the differentially down-regulated genes are known to be important enzymes in the xenobiotic metabolic pathway (Jourenkova-Mironova et al., 1999; Katoh et al., 1999; Park et al., 1997; Sato et al., 1999). These include cytochrome c oxidase subunit Vb (coxVb), gamma-aminobutyraldehyde dehydrogenase, microsomal glutathione S-transferase (GST-II), aldehyde dehydrogenase 7 (ADH7), COX C VIII, ALDH8, EPH2 cytosolic epoxide hydrolase and ALDH10. Further data analysis revealed that other xenobiotic pathway genes, not included in this cluster, were also down-regulated in all five cases, suggesting perhaps a general downregulation of xenobiotic pathway genes during oral cancer development.

The xenobiotic pathway is of importance in the degradative metabolism of both foreign/ native toxic and carcinogenic products. Phase I and II xenobiotic enzymes are two key sequential steps in the metabolism of toxic substances including alcohol and tobacco products. It is interesting to note that most of the five cases of oral cancer were from heavy smokers and drinkers. These data indicate that key regulatory events were altered in the xenobiotic pathways during oral carcinogenesis that may contribute to the increased susceptibility towards carcinogens such as tobacco and alcohol, the two major etiological factors for oral carcinogenesis.

Using Matlab software analysis, 117 transcripts were identified to be differentially expressed between normal and tumor cells. Hierarchical clustering is shown in Figure 5. The distinct clustering of the normal samples from the tumor samples suggests that LCM procured pure, homogenous samples.

Based on the outcome of three analytical methods (GeneChip, SOM and Matlab), ~600 candidate oral cancer genes were identified. Of this comprehensive set, 27 of the differentially expressed genes were identified by all three methods (set forth in Figure 6). Of the 600 candidate genes, 41% were detected at low levels, 1-5 copies per cell.

EXAMPLE IV

Additional Studies

Shillitoe *et al.* and Leethanakul *et al.* have created expression libraries of human oral cancer cell lines and LCM-generated oral cancer tissues (Leethanakul et al., 2000a; Leethanakul et al., 2000b; Shillitoe et al., 2000). Their studies revealed 52 genes to be differentially expressed at >2-fold in at least three of the cancer tissue sets. Of these 52 genes, 26 were present on the Affymetrix GeneChip®. Of these 26 overlapping genes, 18 were called absent (not detectable) in both normal and tumor samples (DP-2/U18422; TIMP-4/U76456; VEGF-C/U43142; FGF3/X14445; FGF5/M37825; FGF6/X63454; IGFBP5/M65062; EGF cripto protein CR1 and 2/M96956; APC/M74088; ERK6/X79483; GDI dissociation protein/U82532; MAP kinase p38/L35253; MKK6/U39657; MEKK3/U78876; Frizzled/L37782; FZD3/U82169; Dishevelled homolog/U46461; Patched homolog/U43148;); one gene showed no difference between normal and tumor tissues (cyclin H/U11791); one gene was upregulated in five out of five tumors (beta1-catenin/X87838); three genes were upregulated in four out of five tumors (thrombospondin2 precursor/L12350; inhibitor of apoptosis protein/U45878; Caspase 5 precursor/U28014); one gene was upregulated in three out of five tumors (MMP-10/X07820); and one gene was downregulated in four out of five tumors (RhoA/L25080). Finally, one gene was upregulated in two tumors, downregulated in two tumors, and called absent in the fifth oral cancer (TRAF2/U78798).

Of the 52 genes, two genes were detected present only through LCM/GeneChip® analysis. They are human SPARC/osteonectin (J03040) and 5T4 oncofetal antigen (Z29083), which are consistently altered in the same manner in all five oral cancers examined.

Of interest is that a number of genes were identified by either LCM/oligonucleotide microarray approach or the LCM/cDNA library approach (Leethanakul et al., 2000a; Leethanakul et al., 2000b; Shillitoe et al., 2000) to be highly expressed/upregulated in oral cancer tissues. These include: ferritin heavy polypeptide I, urokinase plasminogen activator, ATP-binding cassette transporter, interleukin-1 receptor antagonist and keratin 4.

In addition, there are genes that were differentially expressed and detectable in the cell line study (Shillitoe et al., 2000), not in the Head and Neck CGAP (HNCGAP) libraries (Leethanakul et al., 2000a; Leethanakul et al., 2000b), but were detected present in the dataset. Examples of these genes are the collagen type 1 alpha 2 genes and the heat shock protein 70 kD gene. An example of a gene that was not identified by either LCM approach (HNCGAP libraries or the microarray method), but detected present in the cell line filtered cDNA microarray analysis is the transforming growth factor alpha gene, indicating the elevated expression of this gene maybe associated with *in vitro* culturing.

The different outcomes of the various studies are likely reflective of the experimental approaches and methods of analyses. First, by using LCM-generated RNA, contamination of heterogeneous cellular elements is avoided. Second, sample number and the type of microarray used in the respective studies may be relevant to the discrepancies. Third, the stage of the tumor, source and anatomical site of the oral cancers, and handling methods can further result in different gene expression levels. However, LCM-generated RNA, linearly amplified by T7 RNA polymerase and subsequently analyzed by high-density oligonucleotide GeneChip® probe arrays impressively provided for the detection of 39 cellular genes consistently altered in five out of five different paired cases of human oral cancer making these genes useful as classifiers to predict the normal/malignant nature of oral epithelial tissues.

Importantly, the biology associated with these genes could be used to evaluate their role in oral cancer development. A number of these genes are secretory proteins that are upregulated in cancer tissues and could be evaluated as biomarkers of oral malignancy. These include osteonectin, ferritin, cathepsin L, proteoglycan (secretory granule) and oncofetal trophoblast glycoprotein.

All publications and patent applications cited above are incorporated by reference in their entirety for all purposes to the same extent as if each individual publication or patent application were specifically and individually indicated to be so incorporated by reference. Although the present invention has been described in some detail by way of illustration and example for purposes of clarity and understanding, it will be apparent that certain changes and modifications may be practiced within the scope of the appended claims.